**《数据挖掘技术归与应用》课程教学大纲**

**一、课程简介**

|  |  |
| --- | --- |
| **课程中文名** | 数据挖掘技术与应用 |
| **课程英文名** | **Technology and Application of Data Mining**  | **双语授课** | □是 ☑否 |
| **课程代码** | 10122084 | **课程学分** | **4** | **总学时数** | 64（含实践32） |
| **课程类别** | □通识教育课程□公共基础课程☑专业教育课程□综合实践课程□教师教育课程 | **课程性质** | □必修☑选修□其他 | **课程形态** | □线上☑线下□线上线下混合式□社会实践□虚拟仿真实验教学 |
| **考核方式** | ☑闭卷 □开卷 □课程论文 □课程作品 □汇报展示 □报告 ☑课堂表现 ☑阶段性测试 ☑平时作业 ☑其他 （可多选） |
| **开课学院** | 大数据与智能工程学院 | **开课****系(教研室)** | 数据科学与大数据系 |
| **面向专业** | 计算机科学与技术 | **开课学期** | 第5学期 |
| **课程负责人** | 张素兰 | **审核人** | 黄金龙 |
| **先修课程** | 数据结构与算法、Python程序设计 |
| **后续课程** | 企业级应用系统综合实训 |
| **选用教材** | 1.吴建生, 许桂秋. 数据挖掘与机器学习[M], 北京:人民邮电出版社, 2022. |
| **参考书目** | 1.盛胜利, 林海, 李金洪.Python数据挖掘与机器学习实践[M], 北京:机械工业出版社, 2019.2.Jiawei Han, Micheline Kamber, Jian Pei. 数据挖掘概念与技术[M], 北京:机器工业出版社, 2012. |
| **课程资源** | https://www.educoder.net/paths/ijeknthr |
| **课程简介** | 数据挖掘技术与应用是计算机科学与技术专业选修课程，主要介绍数据挖掘基本概念、数据统计分析方法，机器学习框架、分类、回归、聚类、降维、关联规则和协同过滤、离群检测等分析方法。本课程是《数据结构与算法》课程的延伸，在学生数据处理及数据分析能力培养中处于较为重要地位。通过该门课程的学习，不但可以验证、巩固理论知识，加深对数据挖掘算法的理解，而且可以提高学生分析问题、解决实际问题的能力，培养实事求是的科学态度和创新精神，为后续课程的学习和将来就业、科研工作打下基础。 |

**二、课程目标**

**表2-1 课程目标**

|  |  |
| --- | --- |
| **序号** | **具体课程目标** |
| **课程目标 1** | 能够陈述常用数据挖掘算法的原理，习得数据挖掘算法构建方法，阐明机器学习过程模型，能根据实际应用领域数据特征，描述数据预处理、数据分析和数据可视化方法，能描述数据挖掘领域的最新技术和发展趋势。 |
| **课程目标 2** | 具备运用数据挖掘知识、人工智能技术、数据可视化方法及相关分析工具，对计算机复杂工程问题进行研究、分析和解释的能力，具备对分析结果进行归纳和总结能力，逐步形成科学的学习观和方法论。 |
| **课程目标** 3 | 紧跟数据挖掘领域的前沿技术，掌握最新工具的使用方法，逐步养成严谨的科学态度、积极向上的价值观和终身学习的精神，为未来的学习、工作和生活奠定基础。 |

**表2-2 课程目标与毕业要求对应关系（计算机科学与技术）**

| **毕业要求** | **指标点** | **课程目标** |
| --- | --- | --- |
| **1.工程知识：**能够将数学、自然科学、工程基础和计算机科学与技术专业知识用于解决计算机复杂工程问题。【H】 | 1.2 能够针对计算系统及其计算过程，选择或建立适当的设计方案和描述模型。 | 课程目标1 |
| **4.研究：**能够基于科学原理并采用科学方法对计算机复杂工程问题进行研究，设计实验、分析与解释数据，并能通过信息综合得到合理有效结论。【M】 | 4.3 能够对实验结果进行分析和解释，并综合相关信息，给出合理、有效的结论，并对结论进行规范的表述。 | 课程目标2 |
| **5.使用现代工具：**能够针对计算机复杂工程问题，开发、选择与使用恰当的技术、资源、现代工程工具和信息技术工具，包括对计算机复杂工程问题的预测与模拟，并能够理解其局限性。【L】 | 5.3 能够对与计算机软件系统开发相关的复杂工程问题进行预测与模拟所获数据进行分析，并给出解释。 | 课程目标3 |

**三、课程学习内容与方法**

**（一）理论学习内容及要求**

**表3-1 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **课程模块** | **学习内容** | **学习任务** | **课程目标** | **学习重点难点** | **教学方法** | **学时** |
| 1 | 数据挖掘概述 | 1.什么是数据挖掘 | 1.了解数据挖掘的相关概念2.理解监督学习与无监督学习的区别3.查阅相关文献，整理目前数据挖掘的最新的技术 | 1 | **重点：**1.学生对数据挖掘概念有初步的认识2.学生对数据挖掘模式有初步的认识**难点：**3.学生对数据挖掘方法有初步掌握 | 1.讲授法：引导学生掌握数据挖掘的概念。2.专题研讨：促进学生理解不同数据挖掘方法特点。 | 2 |
| 2.数据挖掘的数据类型 | 1 |
| 3.数据挖掘模式 | 1/2 |
| 4.数据挖掘方法 | 1/2 |
| 2 | Pandas数据分析 | 1.统计学与数据挖掘 | 1.了解统计学相关指标在数据挖掘过程中的应用2.掌握Pandas库的数据结构及使用方法3.能够使用Pandas进行初步的数据处理及分析 | 1/2 | **重点：**1.学生对统计学相关指标有初步的了解2.学生掌握Pandas工具的使用**难点：**1.学生掌握使用Pandas进行初步数据处理能力 | 1.讲授法：引导学生掌握Pandas的使用。2.专题研讨：促进学生理解统计分析方法。 | 2 |
| 2.常用统计学指标 | 1/2 |
| 3.Pandas介绍 | 1/2/3 |
| 4.Pandas数据分析 | 1/2/3 |
| 3 | 机器学习基础 | 1.机器学习概述 | 1.掌握机器学习相关概念2.掌握sklearn构建机器学习模型过程3.能够对不同模型使用相应的方法进行评估 | 1 | **重点：**1.学生对机器学习相关概念有初步认识2.学生对机学习框架有初步了解**难点：**1.学生掌握机器学习在数据挖掘中的运用能力 | 1.讲授法：引导学生掌握机器学习的概念。2.专题研讨：促进学生理解机器学习模型构建过程。 | 2 |
| 2.机器学习框架 | 1/2 |
| 3.模型的评判 | 1/2 |
| 4.过拟合问题 | 1/2 |
| 5.KNN算法简单实现及应用 | 1/2/3 |
| 4 | 分类分析方法与应用 | 1.分类分析概述 | 1.掌握分类分析相关概念2.掌握SVM相关概念3.掌握SVM原理4.握概论相关概念5.掌握朴素贝叶斯原理6.能够使用Python语言及构建分类器7.能够对分类模型进行性能评估 | 1 | **重点：**1.学生对分类分析有初步认识2.学生对分类分析算法原理与结果了解**难点：**1.学生掌握对数据初步分类挖掘的操作能力 | 1.讲授法：引导学生掌握机分类算法的原理。2.专题研讨：促进学生掌握分类模型构建及性能评估。 | 6 |
| 2.SVM相关概念 | 1 |
| 3.SVM原理 | 1/2 |
| 4.SVM分类器构建 | 1/2 |
| 5.概率论相关概念 | 1 |
| 6.贝叶斯决策论原理 | 1/2 |
| 7.朴素贝叶斯分类器构建 | 1/2 |
| 8.分类器应用及性能评估 | 1/2/3 |
| 5 | 回归分析方法与应用 | 1.回归分析概述 | 1.掌握回归分析相关概念2.掌握线性回归原理3.掌握多元线性回归原理4.掌握岭回归及Lasso回归原理5.掌握逻辑回归原理 6.能够使用Python语言实现回归模型7.能够对回归模型进行性能评估 | 1 | **重点：**1.学生对回归模型有初步认识2.学生掌握回归算法原理**难点：**1.学生掌握对数据进行回归分析的能力 | 1.讲授法：引导学生掌握各类回归算法原理。2.专题研讨：促进学生掌握回归模型的构建及性能评估。 | 6 |
| 2.线性回归 | 1/2 |
| 3.多元回归 | 1/2 |
| 4.岭回归及Lasso回归 | 1/2 |
| 5.逻辑回归 | 1/2 |
| 6.回归模型应用及性能评估 | 1/2/3 |
| 6 | 聚类分析方法与应用 | 1.聚类分析概述 | 1.掌握聚类分析概念2.掌握K-means聚类算法原理3.掌握DBSCAN聚类算法原理4.掌握层次聚类算法原理5.能够使用Python语言实现聚类模型6.能够对聚类模型进行性能评估 | 1 | **重点：**1.学习对聚类分析有初步的认识2.学生掌握聚类算法的原理 **难点：**1.学生掌握对数据进行聚类分析的能力 | 1.讲授法：引导学生掌握各类聚类算法原理。2.专题研讨：促进学生掌握聚类模型构建及性能评估。 | 4 |
| 2.原型聚类 | 1/2 |
| 3.密度聚类 | 1/2 |
| 4.层次聚类 | 1/2 |
| 5.聚类算法实现 | 1/2/3 |
| 6.聚类算法性能评估 | 1/2/3 |
| 7 | 降维分析方法与应用 | 1.降维分析概述 | 1.掌握降维分析相关概念2.掌握主成分分析算法原理3.掌握流形学习算法原理4.能够使用Python语言实现聚类模型5.能够对聚类模型进行性能评估 | 1 | **重点：**1.学生对降维分析有初步的认识2.学生掌握降维算法原理**难点：**1.学生掌握对数据进行降维分析的能力 | 1.讲授法：引导学生掌握各类降维算法原理。2.专题研讨：促进学生掌握降维模型构建及性能评估。 | 4 |
| 2.主成分分析原理 | 1/2 |
| 3.流形学习原理 | 1/2 |
| 4.降维算法实现 | 1/2/3 |
| 5.降维算法性能评估 | 1/2/3 |
| 8 | 推荐分析方法及应用 | 1.推荐分析概述 | 1.掌握推荐分析相关概念2.掌握关联规则挖掘原理3.掌握Apriori算法原理4.掌握协同过滤算法原理5.能够使用Python语言实现推荐算法6.能够对推荐算法进行性能评估 | 1 | **重点：**1.学生对推荐分析有初步的认识2.学生掌握各类推荐算法原理**难点：**1.学生掌握对数据进行推荐分析的能力 | 1.讲授法：引导学生掌握各类推荐算法原理。2.专题研讨：促进学生掌握推荐模型构建及性能评估。 | 4 |
| 2.关联规则挖掘原理 | 1/2 |
| 3.Apriori算法原理 | 1/2 |
| 4.协同过滤算法原理 | 1/2 |
| 5.推荐算法实现 | 1/2/3 |
| 6.推荐算法性能评估 | 1/2/3 |
| 9 | 离群检测分析方法及应用 | 1.离群检测概述 | 1.掌握离群检测相关概念2.掌握离群检测算法原理3.能够使用Python语言实现离群检测算法4.能够对离群检测算法进行性能评估 | 1 | **重点：**1.学生对离群检测分析有初步的认识2.学生掌握离群检测算法原理**难点：**1.学生掌握对数据进行离群检测的能力 | 1.讲授法：引导学生掌握离群检测算法原理。2.专题研讨：促进学生掌握离群检测模型构建及性能评估。 | 2 |
| 2.离群检测算法原理 | 1/2 |
| 3.离群检测算法实现 | 1/2/3 |
| 4.离群检测算法性能评估 | 1/2/3 |
| 合计 |  |  |  |  |  |  | 32 |

1. **实验学习内容及要求**

**表3-2 课程目标、学习内容和教学方法对应关系**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **项目名称** | **项目来源** | **教学目标（观测点、重难点）** | **学时数** | **项目类型** | **要求** | **每组人数** | **教学方法** | **课程目标** |
| 1 | Pandas数据分析及可视化 | 实验教材 | 1.掌握Pandas工具的使用（重点） | 4 | 验证性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握数据可视化工具及方法（重点） |
| 3.掌握数据预处理及统计分析方法（难点） |
| 2 | 基于sklearn的机器学习模型构建 | 实验教材 | 1.掌握KNN算法原理（重点） | 4 | 验证性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握sklearn下KNN模型构建（重点） |
| 3.使用分类精度、分类误差等对模型进行评估（难点）（难点） |
| 3 | 基于SVM分类器的鸢尾花分类预测 | 实验教材 | 1.掌握SVM分类算法原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握SVM分类模型的构建（重点） |
| 3.使用分类精度、交叉验证精度等对分类模型性能进行评估（难点） |
| 4 | 基于朴素贝叶斯分类器的幸存者预测 | 实验教材 | 1.掌握朴素贝叶斯分类器原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.能够建立朴素贝叶斯分类器，对幸存者进行预测（重点） |
| 3.学习绘制ROC曲线，比较predict和predict\_prob的区别（难点） |
| 5 | 基于回归模型的汽车油耗预测 | 实验教材 | 1.掌握回归算法的原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握回归模型的构建（重点） |
| 3.使用R2、MSE、AIC、BIC、Prob(F-statistic)等指标对对回归模型性能进行评估（难点） |
| 6 | 基于逻辑回归的非线性分类预测 | 实验教材 | 1.掌握逻辑归算法的原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.掌握逻辑回归模型的构建（重点） |
| 3.使用混淆矩阵、精确率、召回率、准确率、f1-score可以量化衡量模型 |
| 4.能够建立决策树分类器，对幸存者进行预测（重点） |
| 5.使用LabelEncoder对离散变量进行编码和简单的缺失值处理方法。（难点） |
| 7 | k-means、层次聚类和DBSCAN聚类对比实验 | 实验教材 | 1.掌握聚类算法原理（重点） | 4 | 设计性实验 | 必做 | 1 | 实验指导 | 1/2/3 |
| 2.能够构建k-means、层次聚类、DBSCAN聚类模型（重点） |
| 3.能够分析kmeans、层次聚类和DBSCAN聚类模型效果差异（难点） |
| 8 | 基于PCA模型的鸢尾花数据可视化 | 实验教材 | 1.掌握PCA算法的原理（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够通过线性组合特征值为“主成分”实现降维（重点） |
| 3.能够使用降维后的数据进行可视化（难点） |
| 9 | 基于FP树的商品推荐模型 | 实验教材 | 1.掌握关联规则挖掘原理（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够将结构化数据转化为“购物篮”数据（重点） |
| 3.能够建立FP关联规则模型，生成规则（难点） |
| 10 | 基于孤立森林的异常点检测 | 实验教材 | 1.掌握离群点检测原理（重点） | 2 | 设计性实验 | 选做 | 1 | 实验指导 | 1/2/3 |
| 2.能够构建孤立森林的异常检测模型（重点） |
| 3.能够分析异常检测结果并对结果进行可视化（难点） |

## 四、课程考核

**（一）考核内容与考核方式**

**表4-1 课程目标、考核内容与考核方式对应关系**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程目标** | **考核内容** | **所属****学习模块/项目** | **考核占比** | **考核方式** |
| 课程目标 1 | 1.数据挖掘概念 | 学习模块1 | 50% | 课堂表现实验项目阶段测验期末考试 |
| 2.机器学习概念 | 学习模块3 |
| 3.分类分析概念 | 学习模块4 |
| 4.SVM  | 学习模块4项目3 |
| 5.贝叶斯决策论 | 学习模块4项目4 |
| 6.回归分析概念 | 学习模块5 |
| 7.线性回归 | 学习模块5项目5 |
| 8.多元回归 | 学习模块5项目5 |
| 9.岭回归及Lasso回归 | 学习模块5项目5 |
| 10.逻辑回归 | 学习模块5项目6 |
| 11.聚类分析概念 | 学习模块6 |
| 12.K-means  | 学习模块6项目7 |
| 13.层次聚类 | 学习模块6项目7 |
| 14.密度聚类 | 学习模块6项目7 |
| 15.降维分析概念 | 学习模块7 |
| 16.主成分分析 | 学习模块7项目8 |
| 17.流形学习 | 学习模块7项目8 |
| 18.推荐分析概念 | 学习模块8 |
| 19.关联规则挖掘 | 学习模块8项目9 |
| 20.协同过滤 | 学习模块8 |
| 21.离群检测 | 学习模块9项目10 |
| 课程目标 2 | 1.数据预测处理 | 学习模块2项目1,3-10 | 31% | 课堂表现实验项目阶段测验期末考试 |
| 2.数据可视化 | 学习模块2项目3-8,10 |
| 3.分类算法实现 | 学习模块4项目3-4 |
| 4.回归算法实现 | 学习模块5项目5-6 |
| 5.聚类算法实现 | 学习模块6项目7 |
| 6.降维算法实现 | 学习模块7项目8 |
| 7.推荐算法实现 | 学习模块8项目9 |
| 8.离群检测算法实现 | 学习模块9项目10 |
| 课程目标 3 | 1.Pandas数据分析 | 学习模块2项目2 | 19% | 课堂表现实验项目阶段测验期末考试 |
| 2.sklearn机器学习模型 | 学习模块3项目3 |
| 3.算法性能评估 | 学习模块4-9项目3-10 |

**表4-2 课程目标与考核方式矩阵关系**

|  |  |  |
| --- | --- | --- |
| 课程目标 | 考核方式 | 考核占比 |
| 平时考核（50%） | 期末考核（50%） |
| 课堂表现（10%） | 阶段测验（15%） | 实验项目（25%） |
| 课程目标1 | 60% | 60% | 40% | 50% | 50%=10%\*60%+25%\*40%+15%\*60%+50%\*50% |
| 课程目标2 | 30% | 20% | 40% | 30% | 31%=10%\*30%+25%\*30%+15%\*20%+50%\*30% |
| 课程目标3 | 10% | 20% | 20% | 20% | 19%=10%\*10%+25%\*30%+15%\*20%+50%\*20% |

## （二）成绩评定

**1. 平时成绩评定**

（1）课堂表现（20分）：通过学生课堂发言、提问及小组讨论情况，评价学生的课程学习态度和参与能力、专业认同感和终身学习意识。

（2）阶段性测验（30分）：以阶段性测验方式进行，评价学生当前的学习情况。

（3）实验项目（50分）：实验课程的实验项目及实验报告完成情况，主要评价学生的实际操作能力。

**2.期末成绩评定**

主要考察学生掌握数据挖掘的基本概念和基础知识情况，对回归、分类、聚类、降维、推荐等智能模型的理解和应用，考核学生进行数据预测处理、数据分析、数据可视化的能力。

**3.总成绩评定**

总成绩（100%）=平时成绩（50%）+期末成绩（50%）

**（三）评分标准**

1.平时成绩

（1）课堂表现：通过学生在课堂上的表现来评价学生对知识的掌握情况及相关能力水平。包括课堂发言、提问、回答问题、测验、练习等，每人每期参加上述课堂活动不得少于两次，少于两次者，课堂表现成绩记为0分（目标1：60%；目标2：30%；目标3：10%)。

（2）阶段测验：根据教学任务，在教学实验系统中安排阶段测验，考察学生对相应章节理论知识的掌握情况，对实际问题分析能力，使用相应工具能力和对结果归纳总结能力。取学生阶段测验平均分，未做测验按“0”分计算，成绩由教学实验系统记录确定（目标1：60%；目标2：20%；目标3：20%)。

（3）实验项目：通过实验教学中的实际案例，考察学生使用相应工具和可视化方法，进行数据回归、分类、聚类、降维、推荐、离群分析的能力，以及对结果进行归纳总结、撰写文档能力。根据学生每个实验项目实验过程的合理性、实验结果正确性、归纳总结的合理性给定成绩（目标1：40%；目标2：40%；目标3：20%)。

2.期末成绩

要考察学生掌握数据挖掘的基本概念和基础知识，对回归、分类、聚类、降维、推荐分析方法的理解与运用，考核学生对特定问题的分析、处理和总结能力。考试方式为考试平台闭卷考试。

## 五、其它说明

本课程大纲依据2023版人才培养方案，由大数据与智能工程学院计算机基础教学部讨论制定，大数据与智能工程学院教学工作委员会审定，教务处审核批准，自2023级开始执行。